On resolving the Savage–Dickey paradox

Jean-Michel Marin and Christian P. Robert

Institut de Mathématiques et Modélisation de Montpellier,
Université Montpellier 2, Case Courrier 51
34095 Montpellier cedex 5, France,
e-mail: jean-michel.marin@univ-montp2.fr

Université Paris-Dauphine, CEREMADE
75775 Paris cedex 16, France,
CREST
92245 Malakoff cedex, France
e-mail: xian@ceremade.dauphine.fr

Abstract: When testing a null hypothesis $H_0 : \theta = \theta_0$ in a Bayesian framework, the Savage–Dickey ratio (Dickey, 1971) is known as a specific representation of the Bayes factor (O'Hagan and Forster, 2004) that only uses the posterior distribution under the alternative hypothesis at θ_0, thus allowing for a plug-in version of this quantity. We demonstrate here that the Savage–Dickey representation is in fact a generic representation of the Bayes factor and that it fundamentally relies on specific measure-theoretic versions of the densities involved in the ratio, instead of being a special identity imposing some mathematically void constraints on the prior distributions. We completely clarify the measure-theoretic foundations of the Savage–Dickey representation as well as of the later generalisation of Verdinelli and Wasserman (1995). We provide furthermore a general framework that produces a converging approximation of the Bayes factor that is unrelated with the approach of Verdinelli and Wasserman (1995) and propose a comparison of this new approximation with their version, as well as with bridge sampling and Chib's approaches.

Keywords and phrases: Bayesian model choice, Bayes factor, bridge sampling, conditional distribution, hypothesis testing, Savage–Dickey ratio, zero measure set.

1. Introduction

From a methodological viewpoint, testing a null hypothesis $H_0 : x \sim f_0(x|\omega_0)$ versus the alternative $H_a : x \sim f_1(x|\omega_1)$ in a Bayesian framework requires the introduction of two prior distributions, $\pi_0(\omega_0)$ and $\pi_1(\omega_1)$, that are defined on the respective parameter spaces. In functional terms, the core object of the Bayesian approach to testing and model choice, the Bayes factor (Jeffreys, 1939, Robert, 2001, O'Hagan and Forster, 2004), is indeed a ratio of two marginal densities taken at the same observation x,

$$B_{01}(x) = \frac{\int \pi_0(\omega_0) f_0(x|\omega_0)\, \mathrm{d}\omega_0}{\int \pi_1(\omega_1) f_1(x|\omega_1)\, \mathrm{d}\omega_1} = \frac{m_0(x)}{m_1(x)} \,.$$

(This quantity $B_{01}(x)$ is then compared to 1 in order to decide about the strength of the support of the data in favour of H_0 or H_a.) It is thus mathematically clearly and uniquely defined, provided both integrals exist and differ from both 0 and ∞. The practical computation of the Bayes factor has generated a large literature on approximative (see, e.g. Chib, 1995, Gelman and Meng, 1998, Chen et al., 2000, Chopin and Robert, 2010), seeking improvements in numerical precision.

The Savage–Dickey (Dickey, 1971) representation of the Bayes factor is primarily known as a special identity that relates the Bayes factor to the posterior distribution which corresponds to the more complex hypothesis. As described in Verdinelli and Wasserman (1995) and Chen et al. (2000, pages 164-165), this representation has practical implications as a basis for simulation methods. However, as stressed in Dickey (1971) and O'Hagan and Forster (2004), the foundation of the Savage–Dickey representation is clearly theoretical.

More specifically, when considering a testing problem with an embedded model, $H_0 : \theta = \theta_0$, and a nuisance parameter ψ, i.e. when ω_1 can be decomposed as $\omega_1 = (\theta, \psi)$ and when $\omega_0 = (\theta_0, \psi)$, for a sampling distribution $f(x|\theta, \psi)$, the plug-in representation

$$B_{01}(x) = \frac{\pi_1(\theta_0|x)}{\pi_1(\theta_0)} \,, \tag{1}$$

with the obvious notations for the marginal distributions

$$\pi_1(\theta) = \int \pi_1(\theta, \psi) \mathrm{d}\psi \quad \text{and} \quad \pi_1(\theta|x) = \int \pi_1(\theta, \psi|x) \mathrm{d}\psi \,,$$

holds under Dickey's (1971) assumption that the conditional prior density of ψ under the alternative model, given $\theta = \theta_0$, $\pi_1(\psi|\theta_0)$, is equal to the prior density under the null hypothesis, $\pi_0(\psi)$,

$$\pi_1(\psi|\theta_0) = \pi_0(\psi) \,. \tag{2}$$

Therefore, Dickey's (1971) identity (1) reduces the Bayes factor to the ratio of the posterior over the prior marginal densities of θ under the alternative model, taken at the tested value θ_0. The Bayes factor is thus expressed as an amount of information brought by the data and this helps in its justification as a model choice tool. (See also Consonni and Veronese, 2008.)

In order to illustrate the Savage–Dickey representation, consider the artificial example of computing the Bayes factor between the models

$$\mathfrak{M}_0: \quad x|\psi \sim \mathcal{N}(\psi, 1), \quad \psi \sim \mathcal{N}(0, 1) \,,$$

and

$$\mathfrak{M}_1: \quad x|\theta, \psi \sim \mathcal{N}(\psi, \theta), \quad \psi|\theta \sim \mathcal{N}(0, \theta), \quad \theta \sim I\mathcal{G}(1, 1) \,,$$

which is equivalent to testing the null hypothesis $H_0: \theta = \theta_0 = 1$ against the alternative $H_1: \theta \neq 1$ when $x|\theta, \psi \sim \mathcal{N}(\psi, \theta)$. In that case, model $\mathfrak{M}_0$ clearly is embedded in model $\mathfrak{M}_1$. We have

$$m_0(x) = \exp\left(-x^2/4\right)/(\sqrt{2}\sqrt{2\pi}) \quad \text{and} \quad m_1(x) = \left(1 + x^2/4\right)^{-3/2} \Gamma(3/2)/(\sqrt{2}\sqrt{2\pi}) \,,$$

and therefore

$$B_{01}(x) = \Gamma(3/2)^{-1} \left(1 + x^2/4\right)^{3/2} \exp\left(-x^2/4\right) \,.$$

Dickey's assumption (2) on the prior densities is satisfied, since

$$\pi_1(\psi|\theta_0) = \frac{1}{\sqrt{2\pi}} \exp\left(-\psi^2/2\right) = \pi_0(\psi) \,.$$

Therefore, since

$$\pi_1(\theta) = \theta^{-2} \exp\left(-\theta^{-1}\right) \,, \quad \pi_1(\theta_0) = \exp(-1) \,,$$

and

$$\pi_1(\theta|x) = \Gamma(3/2)^{-1} \left(1 + x^2/4\right)^{3/2} \theta^{-5/2} \exp\left(-\theta^{-1}\left(1 + x^2/4\right)\right) \mathbb{I}_{\theta>0} \,,$$
$$\pi_1(\theta_0|x) = \Gamma(3/2)^{-1} \left(1 + x^2/4\right)^{3/2} \exp\left(-\left(1 + x^2/4\right)\right) \,,$$

we clearly recover the Savage–Dickey representation

$$B_{01}(x) = \Gamma(3/2)^{-1} \left(1 + x^2/4\right)^{3/2} \exp\left(-x^2/4\right) = \pi_1(\theta_0|x)/\pi_1(\theta_0) \,.$$

While the difficulty with the representation (1) is usually addressed in terms of computational aspects, given that $\pi_1(\theta|x)$ is rarely available in closed form, we argue in the current paper that the Savage–Dickey representation faces challenges of a deeper nature that led us to consider it a 'paradox'. First, by considering both prior and posterior marginal distributions of θ uniquely *under the alternative model,* (1) seems to indicate that the posterior probability of the null hypothesis $H_0: \theta = \theta_0$ is contained within the alternative hypothesis posterior distribution, even though the set of (θ, ψ)'s such that $\theta = \theta_0$ has a zero probability under this alternative distribution. Second, as explained in Section 2, an even more fundamental difficulty with assumption (2) is that it is meaningless when examined (as it should) within the mathematical axioms of measure theory.

Having stated those mathematical difficulties with the Savage–Dickey representation, we proceed to show in Section 3 that similar identities hold under no constraint on the prior distributions. In Section 3, we derive computational algorithms that exploit these representations to approximate the Bayes factor, in an approach that differs from the earlier solution of Verdinelli and Wasserman (1995). The paper concludes with an illustration in the setting of variable selection within a probit model.

2. A measure-theoretic paradox

When considering a standard probabilistic setting where the dominating measure on the parameter space is the Lebesgue measure, rather than a counting measure, the conditional density $\pi_1(\psi|\theta)$ is rigorously (Billingsley, 1986) defined as the density of the conditional probability distribution or, equivalently, by the condition that

$$\mathbb{P}((\theta, \psi) \in A_1 \times A_2) = \int_{A_1} \int_{A_2} \pi_1(\psi|\theta) \, \mathrm{d}\psi \, \pi_1(\theta) \, \mathrm{d}\theta = \int_{A_1 \times A_2} \pi_1(\theta, \psi) \mathrm{d}\psi \, \mathrm{d}\theta \,,$$

for all measurable sets $A_1 \times A_2$, when $\pi_1(\theta)$ is the associated marginal density of θ. Therefore, this identity points out the well-known fact that the conditional density function $\pi_1(\psi|\theta)$ is defined up to a set of measure zero both in ψ for *every* value of θ *and* in θ. This implies that changing arbitrarily the value of the *function* $\pi_1(\cdot|\theta)$ for a negligible collection of values of θ does not impact the properties of the conditional distribution.

In the setting where the Savage–Dickey representation is advocated, the value θ_0 to be tested is not determined from the observations but it is instead given in advance since this is a testing problem. Therefore the density function

$$\pi_1(\psi|\theta_0)$$

may be chosen in a *completely arbitrary* manner and there is no possible reason for a unique representation of $\pi_1(\psi|\theta_0)$ that can be found within measure theory. This implies that there always is a version of the conditional density $\pi_1(\psi|\theta_0)$ such that Dickey's (1971) condition (2) is satisfied—as well as, conversely, there are an infinity of versions for which it is *not* satisfied—. As a result, from a mathematical perspective, condition (2) cannot be seen as an *assumption* on the prior π_1 without further conditions, contrary to what is stated in the original Dickey (1971) and later in O'Hagan and Forster (2004), Consonni and Veronese (2008) and Wetzels et al. (2010). This difficulty is the first part of what we call the *Savage–Dickey paradox*, namely that, as stated, the representation (1) relies on a mathematically void constraint on the prior distribution. In the specific case of the artificial example introduced above, the choice of the conditional density $\pi_1(\psi|\theta_0)$ is therefore arbitrary: if we pick for this density the density of the $\mathcal{N}(0, 1)$ distribution, there is agreement between $\pi_1(\psi|\theta_0)$ and $\pi_0(\psi)$, while, if we select instead the function $\exp(+\psi^2/2)$, which is not a density, there is no agreement in the sense of condition (2). The paradox is that this disagreement has no consequence whatsoever in the Savage–Dickey representation.

The second part of the Savage–Dickey paradox is that the representation (1) is solely valid for a specific and unique choice of a version of the density for both the conditional density $\pi_1(\psi|\theta_0)$ and the joint density $\pi_1(\theta_0, \psi)$. When looking at the derivation of (1), the choices of some specific versions of those densities are indeed noteworthy: in the following development,

$$
\begin{aligned}
B_{01}(x) &= \frac{\int \pi_0(\psi) f(x|\theta_0, \psi) \, \mathrm{d}\psi}{\int \pi_1(\theta, \psi) f(x|\theta, \psi) \, \mathrm{d}\psi \mathrm{d}\theta} &&\text{[by definition]}\\
&= \frac{\int \pi_1(\psi|\theta_0) f(x|\theta_0, \psi) \, \mathrm{d}\psi \, \pi_1(\theta_0)}{\int \pi_1(\theta, \psi) f(x|\theta, \psi) \, \mathrm{d}\psi \mathrm{d}\theta \, \pi_1(\theta_0)} &&\text{[using a specific version of } \pi_1(\psi|\theta_0)]\\
&= \frac{\int \pi_1(\theta_0, \psi) f(x|\theta_0, \psi) \, \mathrm{d}\psi}{m_1(x) \pi_1(\theta_0)} &&\text{[using a specific version of } \pi_1(\theta_0, \psi)]\\
&= \frac{\pi_1(\theta_0|x)}{\pi_1(\theta_0)} \,, &&\text{[using a specific version of } \pi_1(\theta_0|x)]
\end{aligned}
$$

the second equality depends on a specific choice of the version of $\pi_1(\psi|\theta_0)$ but not on the choice of the version of $\pi_1(\theta_0)$, while the third equality depends on a specific choice of the version of $\pi_1(\psi, \theta_0)$ as equal to $\pi_0(\psi)\pi_1(\theta_0)$, thus related to the choice of the version of $\pi_1(\theta_0)$. The last equality leading to the Savage–Dickey representation relies on the choice of a specific version of $\pi_1(\theta_0|x)$ as well, namely that the constraint

$$\frac{\pi_1(\theta_0|x)}{\pi_1(\theta_0)} = \frac{\int \pi_0(\psi) f(x|\theta_0, \psi) \, \mathrm{d}\psi}{m_1(x)}$$

3

holds, where the right hand side is equal to the Bayes factor $B_{01}(x)$ and is therefore independent from the version. This rigorous analysis implies that the Savage–Dickey representation is tautological, due to the availability of a version of the posterior density that makes it hold.

As an illustration, consider once again the artificial example above. As already stressed, the value to be tested $\theta_0 = 1$ is set prior to the experiment. Thus, without modifying either the prior distribution under model $\mathfrak{M}_1$ or the marginal posterior distribution of the parameter θ under model $\mathfrak{M}_1$, and in a completely rigorous measure-theoretic framework, we can select

$$\pi_1(\theta_0) = 100 = \pi_1(\theta_0|x).$$

For that choice, we obtain

$$\pi_1(\theta_0|x)/\pi_1(\theta_0) = 1 \neq B_{01}(x) = \Gamma(3/2)^{-1} \left(1 + x^2/4\right)^{3/2} \exp\left(-x^2/4\right).$$

Hence, for this specific choice of the densities, the Savage–Dickey representation does not hold.

Verdinelli and Wasserman (1995) have proposed a generalisation of the Savage–Dickey density ratio when the constraint (2) on the prior densities is not verified (we stress again that this is a mathematically void constraint on the respective prior distributions). Verdinelli and Wasserman (1995) state that

$$\begin{aligned}
B_{01}(x) &= \frac{\int \pi_0(\psi) f(x|\theta_0, \psi) \, d\psi}{m_1(x)} && \text{[by definition]} \\
&= \pi_1(\theta_0|x) \frac{\int \pi_0(\psi) f(x|\theta_0, \psi) \, d\psi}{m_1(x)\pi_1(\theta_0|x)} && \text{[for any version of } \pi_1(\theta_0|x)] \\
&= \pi_1(\theta_0|x) \int \frac{\pi_0(\psi) f(x|\theta_0, \psi)}{m_1(x)\pi_1(\theta_0|x)} \frac{\pi_1(\psi|\theta_0)}{\pi_1(\psi|\theta_0)} \, d\psi && \text{[for any version of } \pi_1(\psi|\theta_0)] \\
&= \pi_1(\theta_0|x) \int \frac{\pi_0(\psi)}{\pi_1(\psi|\theta_0)} \frac{f(x|\theta_0, \psi)\pi_1(\psi|\theta_0) \, d\psi}{m_1(x)\pi_1(\theta_0|x)} \frac{\pi_1(\theta_0)}{\pi_1(\theta_0)} && \text{[for any version of } \pi_1(\theta_0)] \\
&= \frac{\pi_1(\theta_0|x)}{\pi_1(\theta_0)} \int \frac{\pi_0(\psi)}{\pi_1(\psi|\theta_0)} \pi_1(\psi|\theta_0, x) \, d\psi && \text{[for a specific version of } \pi_1(\psi|\theta_0, x)] \\
&= \frac{\pi_1(\theta_0|x)}{\pi_1(\theta_0)} \mathbb{E}^{\pi_1(\psi|x,\theta_0)} \left[\frac{\pi_0(\psi)}{\pi_1(\psi|\theta_0)}\right].
\end{aligned}$$

This representation of Verdinelli and Wasserman (1995) therefore remains valid for any choice of versions for $\pi_1(\theta_0|x)$, $\pi_1(\theta_0)$, $\pi_1(\psi|\theta_0)$, provided the conditional density $\pi_1(\psi|\theta_0, x)$ is defined by

$$\pi_1(\psi|\theta_0, x) = \frac{f(x|\theta_0, \psi)\pi_1(\psi|\theta_0)\pi_1(\theta_0)}{m_1(x)\pi_1(\theta_0|x)},$$

which obviously means that the Verdinelli–Wasserman representation

$$B_{01}(x) = \frac{\pi_1(\theta_0|x)}{\pi_1(\theta_0)} \mathbb{E}^{\pi_1(\psi|x,\theta_0)} \left[\frac{\pi_0(\psi)}{\pi_1(\psi|\theta_0)}\right] \tag{3}$$

is dependent on the choice of a version of $\pi_1(\theta_0)$.

We now establish that an alternative representation of the Bayes factor is available and can be exploited towards approximation purposes. When considering the Bayes factor

$$B_{01}(x) = \frac{\int \pi_0(\psi) f(x|\theta_0, \psi) \, d\psi}{\int \pi_1(\theta, \psi) f(x|\theta, \psi) \, d\psi d\theta} \frac{\pi_1(\theta_0)}{\pi_1(\theta_0)},$$

where the right hand side obviously is independent of the choice of the version of $\pi_1(\theta_0)$, the numerator can be seen as involving a specific version in $\theta = \theta_0$ of the marginal posterior density

$$\tilde{\pi}_1(\theta|x) \propto \int \pi_0(\psi) f(x|\theta, \psi) \, d\psi \, \pi_1(\theta),$$

4

which is associated with the alternative prior $\tilde{\pi}_1(\theta, \psi) = \pi_1(\theta)\pi_0(\psi)$. Indeed, this density $\tilde{\pi}_1(\theta|x)$ appears as the marginal posterior density of the posterior distribution defined by the density

$$\tilde{\pi}_1(\theta, \psi|x) = \frac{\pi_0(\psi)\pi_1(\theta)f(x|\theta, \psi)}{\tilde{m}_1(x)} ,$$

where $\tilde{m}_1(x)$ is the proper normalising constant of the joint posterior density. In order to guarantee a Savage–Dickey-like representation of the Bayes factor, the appropriate version of the marginal posterior density in $\theta = \theta_0$, $\tilde{\pi}_1(\theta_0|x)$, is obtained by imposing

$$\frac{\tilde{\pi}_1(\theta_0|x)}{\pi_0(\theta_0)} = \frac{\int \pi_0(\psi)f(x|\theta_0, \psi)\, d\psi}{\tilde{m}_1(x)} , \tag{4}$$

where, once again, the right hand side of the equation is uniquely defined. This constraint amounts to imposing that Bayes' theorem holds in $\theta = \theta_0$ instead of almost everywhere (and thus not necessarily in $\theta = \theta_0$). It then leads to the alternative representation

$$B_{01}(x) = \frac{\tilde{\pi}_1(\theta_0|x)}{\pi_1(\theta_0)} \frac{\tilde{m}_1(x)}{m_1(x)} ,$$

which holds for any value chosen for $\pi_1(\theta_0)$ provided condition (4) applies.

This new representation may seem to be only formal, since both $m_1(x)$ and $\tilde{m}_1(x)$ are usually unavailable in closed form, but we can take advantage of the fact that the bridge sampling identity of Torrie and Valleau (1977) (see also Gelman and Meng, 1998) gives an unbiased estimator of $\tilde{m}_1(x)/m_1(x)$ since

$$\mathbb{E}^{\pi_1(\theta, \psi|x)}\left[\frac{\pi_0(\psi)\pi_1(\theta)f(x|\theta, \psi)}{\pi_1(\theta, \psi)f(x|\theta, \psi)}\right] = \mathbb{E}^{\pi_1(\theta, \psi|x)}\left[\frac{\pi_0(\psi)}{\pi_1(\psi|\theta)}\right] = \frac{\tilde{m}_1(x)}{m_1(x)} .$$

In conclusion, we obtain the representation

$$B_{01}(x) = \frac{\tilde{\pi}_1(\theta_0|x)}{\pi_1(\theta_0)} \mathbb{E}^{\pi_1(\theta, \psi|x)}\left[\frac{\pi_0(\psi)}{\pi_1(\psi|\theta)}\right] , \tag{5}$$

whose expectation part is uniquely defined (in that it does not depend on the choice of a version of the densities involved therein), while the first ratio must satisfy condition (4). We further note that this representation clearly differs from Verdinelli and Wasserman's (1995) representation:

$$B_{01}(x) = \frac{\pi_1(\theta_0|x)}{\pi_1(\theta_0)} \mathbb{E}^{\pi_1(\psi|x, \theta_0)}\left[\frac{\pi_0(\psi)}{\pi_1(\psi|\theta_0)}\right] , \tag{6}$$

since (6) uses a specific version of the marginal posterior density on θ in θ_0, as well as a specific version of the full conditional posterior density of ψ given θ_0

3. Computational solutions

In this Section, we consider the computational implications of the above representation in the specific case of latent variable models, namely under the practical possibility of a data completion by a latent variable z such that

$$f(x|\theta, \psi) = \int f(x|\theta, \psi, z)f(z|\theta, \psi)\, dz$$

when $\pi_1(\theta|x, \psi, z) \propto \pi_1(\theta)f(x|\theta, \psi, z)$ is available in closed form, including the normalising constant.

We first consider a computational solution that approximates the Bayes factor based on our novel representation (5). Given a sample $(\bar{\theta}^{(1)}, \bar{\psi}^{(1)}, \bar{z}^{(1)}), \ldots, (\bar{\theta}^{(T)}, \bar{\psi}^{(T)}, \bar{z}^{(T)})$ simulated from (or converging to) the augmented posterior distribution $\tilde{\pi}_1(\theta, \psi, z|x)$, the sequence

$$\frac{1}{T}\sum_{t=1}^{T}\tilde{\pi}_1(\theta_0|x, \bar{z}^{(t)}, \bar{\psi}^{(t)})$$

5

converges to $\tilde{\pi}_1(\theta_0|x)$ in T under the following constraint on the selected version of $\tilde{\pi}_1(\theta_0|x, z, \psi)$ used therein:

$$\frac{\tilde{\pi}_1(\theta_0|x, z, \psi)}{\pi_1(\theta_0)} = \frac{f(x, z|\theta_0, \psi)}{\int f(x, z|\theta, \psi)\pi_1(\theta)\,\mathrm{d}\theta}.$$

which again amounts to imposing that Bayes' theorem holds in $\theta = \theta_0$ for $\tilde{\pi}_1(\theta|x, z, \psi)$ rather than almost everywhere. (Note once more that the right hand side is uniquely defined, i.e. that it does not depend on a specific version.) Therefore, provided iid or MCMC simulations from the joint target $\tilde{\pi}_1(\theta, \psi, z|x)$ are available, the converging approximation to the Bayes factor $B_{01}(x)$ is then

$$\frac{1}{T}\sum_{t=1}^{T}\frac{\tilde{\pi}_1(\theta_0|x, \bar{z}^{(t)}, \bar{\psi}^{(t)})}{\pi_1(\theta_0)}\frac{\tilde{m}_1(x)}{m_1(x)}.$$

(We stress that the simulated sample is produced for the artificial target $\tilde{\pi}_1(\theta, \psi, z|x)$ rather than the true posterior $\pi_1(\theta, \psi, z|x)$ if $\tilde{\pi}_1(\theta, \psi) \neq \pi_1(\theta, \psi)$.) Moreover, if $(\theta^{(1)}, \psi^{(1)}), \ldots, (\theta^{(T)}, \psi^{(T)})$ is a sample independently simulated from (or converging to) $\pi_1(\theta, \psi|x)$, then

$$\frac{1}{T}\sum_{t=1}^{T}\frac{\pi_0(\psi^{(t)})}{\pi_1(\psi^{(t)}|\theta^{(t)})}$$

is a convergent and unbiased estimator of $\tilde{m}_1(x)/m_1(x)$. Therefore, the computational solution associated to our representation (5) of $B_{01}(x)$ leads to the following unbiased estimator of the Bayes factor:

$$\widehat{B_{01}}^{\mathrm{MR}}(x) = \frac{1}{T}\sum_{t=1}^{T}\frac{\tilde{\pi}_1(\theta_0|x, \bar{z}^{(t)}, \bar{\psi}^{(t)})}{\pi_1(\theta_0)}\frac{1}{T}\sum_{t=1}^{T}\frac{\pi_0(\psi^{(t)})}{\pi_1(\psi^{(t)}|\theta^{(t)})}. \tag{7}$$

Note that

$$\mathbb{E}^{\tilde{\pi}_1(\theta, \psi|x)}\left[\frac{\pi_1(\theta, \psi)f(x|\theta, \psi)}{\pi_0(\psi)\pi_1(\theta)f(x|\theta, \psi)}\right] = \mathbb{E}^{\tilde{\pi}_1(\theta, \psi|x)}\left[\frac{\pi_1(\psi|\theta)}{\pi_0(\psi)}\right] = \frac{m_1(x)}{\tilde{m}_1(x)}$$

implies that

$$T\bigg/\sum_{t=1}^{T}\frac{\pi_1(\bar{\psi}^{(t)}|\theta^{(t)})}{\pi_0(\bar{\psi}^{(t)})}$$

is another convergent (if biased) estimator of $\tilde{m}_1(x)/m_1(x)$. The availability of two estimates of the ratio $\tilde{m}_1(x)/m_1(x)$ is a major bonus from a computational point of view since the comparison of both estimators may allow for the detection of infinite variance estimators, as well as for coherence of the approximations. The first approach requires two simulation sequences, one from $\tilde{\pi}_1(\theta, \psi|x)$ and one from $\pi_1(\theta, \psi|x)$, but this is a void constraint in that, if H_0 is rejected, a sample from the alternative hypothesis posterior will be required no matter what. Although we do not pursue this possibility in the current paper, note that a comparison of the different representations (including Verdinelli and Wasserman's, 1995, as exposed below) could be conducted by expressing them in the bridge sampling formalism (Gelman and Meng, 1998).

We now consider a computational solution that approximates the Bayes factor and is based on Verdinelli and Wasserman (1995)'s representation (6). Given a sample $(\theta^{(1)}, \psi^{(1)}, z^{(1)}), \ldots, (\theta^{(T)}, \psi^{(T)}, z^{(T)})$ simulated from (or converging to) $\pi_1(\theta, \psi, z|x)$, the sequence

$$\frac{1}{T}\sum_{t=1}^{T}\pi_1(\theta_0|x, z^{(t)}, \psi^{(t)})$$

converges to $\pi_1(\theta_0|x)$ under the following constraint on the selected version of $\pi_1(\theta_0|x, z, \psi)$ used there:

$$\frac{\pi_1(\theta_0|x, z, \psi)}{\pi_1(\theta_0)} = \frac{f(x, z|\theta_0, \psi)}{\int f(x, z|\theta, \psi)\pi_1(\theta)\,\mathrm{d}\theta}.$$

Moreover, if $\left(\tilde{\psi}^{(1)}, \tilde{z}^{(1)}\right), \ldots, \left(\tilde{\psi}^{(T)}, \tilde{z}^{(T)}\right)$ is a sample generated from (or converging to) $\pi_1(\psi, z|x, \theta_0)$, the sequence

$$\frac{1}{T} \sum_{t=1}^{T} \frac{\pi_0(\tilde{\psi}^{(t)})}{\pi_1(\tilde{\psi}^{(t)}|\theta_0)}$$

is converging to

$$\mathbb{E}^{\pi_1(\psi|x,\theta_0)} \left[\frac{\pi_0(\psi)}{\pi_1(\psi|\theta_0)} \right]$$

under the constraint

$$\pi_1(\psi, z|\theta_0, x) \propto f(x, z|\theta_0, \psi)\pi_1(\psi|\theta_0).$$

Therefore, the computational solution associated to the Verdinelli and Wasserman (1995)'s representation of $B_{01}(x)$ (6) leads to the following unbiased estimator of the Bayes factor:

$$\widehat{B_{01}}^{\text{VW}}(x) = \frac{1}{T} \sum_{t=1}^{T} \frac{\pi_1(\theta_0|x, z^{(t)}, \psi^{(t)})}{\pi_1(\theta_0)} \frac{1}{T} \sum_{t=1}^{T} \frac{\pi_0(\tilde{\psi}^{(t)})}{\pi_1(\tilde{\psi}^{(t)}|\theta_0)}. \tag{8}$$

Although, at first sight, the approximations (7) and (8) may look very similar, the simulated sequences used in both approximations differ: the first average involves simulations from $\tilde{\pi}_1(\theta, \psi, z|x)$ and from $\pi_1(\theta, \psi, z|x)$, respectively, while the second average relies on simulations from $\pi_1(\theta, \psi, z|x)$ and from $\pi_1(\psi, z|x, \theta_0)$, respectively.

4. An illustration

Although our purpose in this note is far from advancing the superiority of the Savage–Dickey type representations for Bayes factor approximation, given the wealth of available solutions for embedded models (Chen et al., 2000, Marin and Robert, 2010), we briefly consider an example where both Verdinelli and Wasserman's (1995) and our proposal apply. The model is the Bayesian posterior distribution of the regression coefficients of a probit model, following the prior modelling adopted in Marin and Robert (2007) that extends Zellner's (1971) g-prior to generalised linear models. We take as data the Pima Indian diabetes study available in R (R Development Core Team, 2008) dataset with 332 women registered and build a probit model predicting the presence of diabetes from three predictors, the glucose concentration, the diastolic blood pressure and the diabetes pedigree function, assessing the impact of the diabetes pedigree function, i.e. testing the nullity of the coefficient θ associated to this variable. For more details on the statistical and computational issues, see Marin and Robert (2010) since this paper relies on the Pima Indian probit model as benchmark.

This probit model is a natural setting for completion by a truncated normal latent variable (Albert and Chib, 1993). We can thus easily implement a Gibbs sampler to produce output from all the posterior distributions considered in the previous Section. Besides, in that case, the conditional distribution $\pi_1(\theta|x, \psi, z)$ is a normal distribution with closed form parameters. It is therefore straightforward to compute the unbiased estimators (7) and (8). Figure 1 compares the variation of this approximation with other standard solutions covered in Marin and Robert (2010) for the same example, namely the regular importance sampling approximation based on the MLE asymptotic distribution, Chib's version based on the same completion, and a bridge sampling (Gelman and Meng, 1998) solution completing $\pi_0(\cdot)$ with the full conditional being derived from the conditional MLE asymptotic distribution. The boxplots are all based on 100 replicates of $T = 20,000$ simulations. While the estimators (7) and (8) are not as accurate as Chib's version and as the importance sampler in this specific case, their variabilities remain at a reasonable order and are very comparable. The R code and the reformated datasets used in this Section are available at the following address: http://www.math.univ-montp2.fr/~marin/savage/dickey.html.
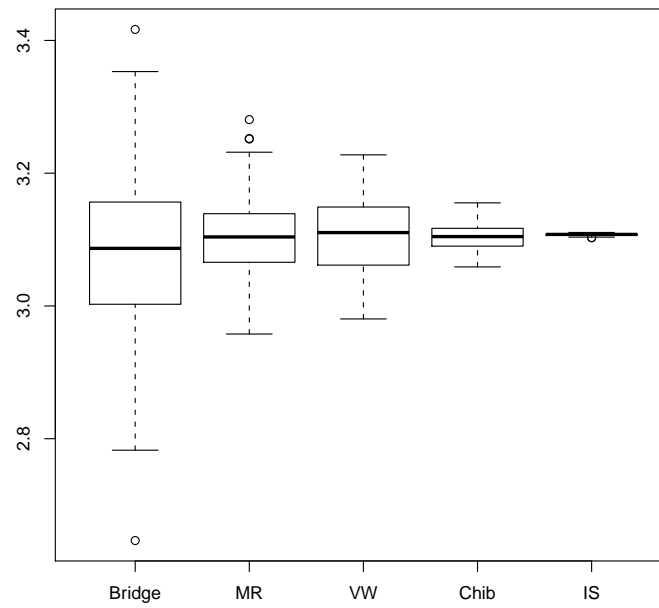
FIG 1. *Comparison of the variabilities of five approximations of the Bayes factor evaluating the impact of the diabetes pedigree covariate upon the occurrence of diabetes in the Pima Indian population, based on a probit modelling. The boxplots are based on 100 replicas and the Savage–Dickey representation proposed in the current paper is denoted by MR, while Verdinelli and Wasserman's (1995) version is denoted by VW.*

References

ALBERT, J. and CHIB, S. (1993). Bayesian analysis of binary and polychotomous response data. *J. American Statist. Assoc.*, **88** 669–679.

BILLINGSLEY, P. (1986). *Probability and Measure.* 2nd ed. John Wiley, New York.

CHEN, M., SHAO, Q. and IBRAHIM, J. (2000). *Monte Carlo Methods in Bayesian Computation.* Springer-Verlag, New York.

CHIB, S. (1995). Marginal likelihood from the Gibbs output. *J. American Statist. Assoc.*, **90** 1313–1321.

CHOPIN, N. and ROBERT, C. (2010). Properties of evidence. *Biometrika.* To appear.

CONSONNI, G. and VERONESE, P. (2008). Compatibility of prior specifications across linear models. *Statist. Science*, **23** 332–353.

DICKEY, J. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. *Ann. Mathemat. Statist.*, **42** 204–223.

GELMAN, A. and MENG, X. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statist. Science*, **13** 163–185.

JEFFREYS, H. (1939). *Theory of Probability.* 1st ed. The Clarendon Press, Oxford.

MARIN, J. and ROBERT, C. (2010). Importance sampling methods for Bayesian discrimination between embedded models. In *Frontiers of Statistical Decision Making and Bayesian Analysis* (M.-H. Chen, D. Dey, P. Müller, D. Sun and K. Ye, eds.). Springer-Verlag, New York. To appear, see arXiv:0910.2325.

MARIN, J.-M. and ROBERT, C. (2007). *Bayesian Core.* Springer-Verlag, New York.

O'HAGAN, A. and FORSTER, J. (2004). *Kendall's advanced theory of Statistics: Bayesian inference.* Arnold, London.

R DEVELOPMENT CORE TEAM (2008). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

ROBERT, C. (2001). *The Bayesian Choice.* 2nd ed. Springer-Verlag, New York.

TORRIE, G. and VALLEAU, J. (1977). Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comp. Phys.*, **23** 187–199.

VERDINELLI, I. and WASSERMAN, L. (1995). Computing Bayes factors using a generalization of the Savage–Dickey density ratio. *J. American Statist. Assoc.*, **90** 614–618.

WETZELS, R., GRASMAN, R. and WAGENMAKERS, E.-J. (2010). An encompassing prior generalization of the Savage-Dickey density ratio. *Comput. Statist. Data Anal.*, **54** 2094–2102.

ZELLNER, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distribution regression using Bayesian variable selection. In *Bayesian inference and decision techniques: Essays in Honor of Bruno de Finetti.* North-Holland / Elsevier, 233–243.